Gesture Control of HMM-Based Singing Voice Synthesis

Gesture Control of HMM-Based Singing Voice Synthesis

Christophe Veaux[1], Maria Astrinaki[2], Keiichiro Oura[3], Robert A.J Clark[1], Junichi Yamagishi[1]

[1] Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] TCTS Lab, University of Mons, Belgium
[3] Departement of Computer Science, Nagoya Institute of Technology, Japan
E-mail: {cveaux, jyamagis}@inf.ed.ac.uk

Abstract

The flexibility of statistical parametric speech synthesis has recently led to the development of interactive speech synthesis systems where different aspects of the voice output can be continuously controlled. The demonstration presented in this paper is based on MAGE/pHTS, a real-time synthesis system developed at Mons University. This system enhances the controllability and the reactivity of HTS by enabling the generation of the speech parameters on the fly. This demonstration gives an illustration of the new possibilities offered by this approach in terms of interaction. A kinect sensor is used to follow the gestures and body posture of the user and these physical parameters are mapped to the prosodic parameters of an HMM-based singing voice model. In this way, the user can directly control various aspect of the singing voice such as the vibrato, the fundamental frequency or the duration. An avatar is used to encourage and facilitate the user interaction.

Index Terms: Performative Speech Synthesis, Mage, Singing Voice Synthesis

1. Introduction

In typical text-to-speech systems, the conversion between text and generated voice occurs at the sentence level. This precludes any form of real-time interaction and control of the output voice, narrowing down the set of potential application of these systems. However, HMM-based speech synthesis [1] brings a new level of flexibility in the generation of the speech parameters. It allows for instance to change the characteristics of a voice [2] or to interpolate between voice models [3] and therefore constitutes a promising framework to explore interactive applications of speech synthesis. Recently, the University of Mons has introduced a modified version of HTS [4], called performative HTS or pHTS [5]. With this new system, the HMM models and the speech parameters can be modified on the fly, enabling a real-time control of the voice output. The demonstration prototype presented in this paper gives an illustration of the possibilities offered by this approach. It uses the pHTS engine in order to modify in real-time the prosodic parameters of an HMM-based singing voice model [6]. The prosodic parameters are mapped to the gestures and body posture of the user tracked by a Microsoft Kinect sensor [7]. This kind of gesture control is particularly well suited for real-time interaction where several parameters can be controlled at the same time. It also provides a physical experience of several dimensions of the singing voice prosody. The section 2 of this paper presents a short overview of pHTS and its real-time software environment MAGE. The prototype used for gesture-controlled singing voice synthesis will be detailed in section 3.

2. Mage/pHTS

The pHTS engine relies on a series of modifications of HTS, enabling a much more reactive control of speech output. The main modification is the reduction of the phonetic context used for the generation of the speech parameters trajectories. The speech parameters are generated using a 2-label sliding window and the corresponding speech sound can be synthesized right away as shown in Figure 1. Within appropriate real-time audio software architecture, it means that the sound can be synthesized on the fly. As a result, any kind of modification performed on the models during the generation of the speech parameters has an impact on the corresponding speech sound with a delay of only one label.

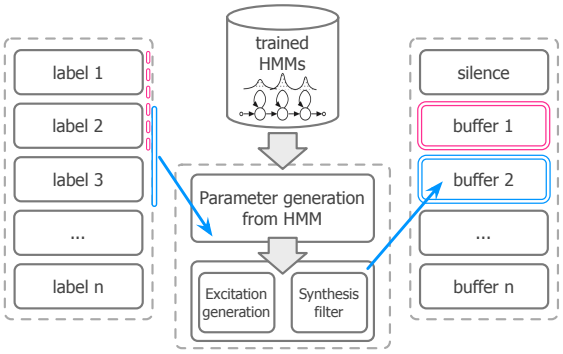

Figure 1: pHTS synthesis using 2-label sliding window to generate the speech parameters.

MAGE is the software umbrella that provides the appropriate real-time audio architecture for the pHTS engine. It also provides a user-friendly API and can be embedded in the PureData programming environment as an external.

3. Gesture control of singing voice synthesis

The MAGE real-time library has already been used in various prototypes exploring how HMM-based speech synthesis can be controlled by gestures. The concept of gesture is here considered in a very large sense, including finger gestures [8] as well as mouth expressions [9]. We present here a new prototype for the gesture control of HMM-based singing voice synthesis. This prototype is based on the following main components, as shown in Figure 2:

- Kinect sensor and its Natural User Interface (NUI) library.
- PureData, real-time programming environment.
- Mage/pHTS, real-time synthesis library.
- Animata, real-time avatar animation software.

The NUI runtime library makes it possible to recognize and track the skeleton joints of a user in front of the Kinect sensor. The skeleton joints coordinates are send to PureData through Open Sound Control (OSC) messages. A series of PureData patches convert the absolute joints coordinates into relative positions and normalizes them by the body size of the user. Finally the gesture descriptors are used to modify the prosodic parameters of the HMM-based singing voice through the MAGE API. The skeleton joints coordinates are also sent to the real-time avatar animation software.

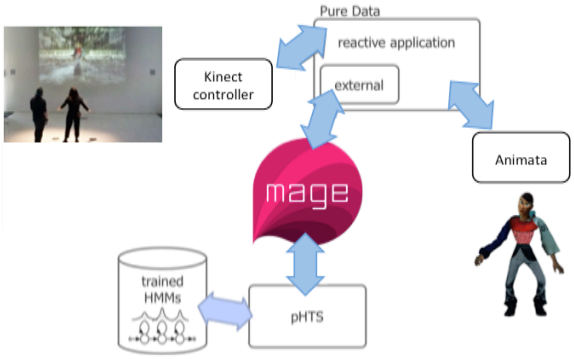

Figure 2: Layout of the gesture-controlled singing synthesis.

The MAGE API was slightly modified in order to enable the control of the following prosodic parameters:

- Vibrato
- Fundamental frequency F0
- Singing speed
- Vocal tract length VTL
- Voicing / Breathiness

The voicing / breathiness parameter was initially introduced for the real-time control of speech synthesis and is not used as a control for the singing synthesis. The mapping between the other prosodic parameters and the hands positions is illustrated in Figure 3.

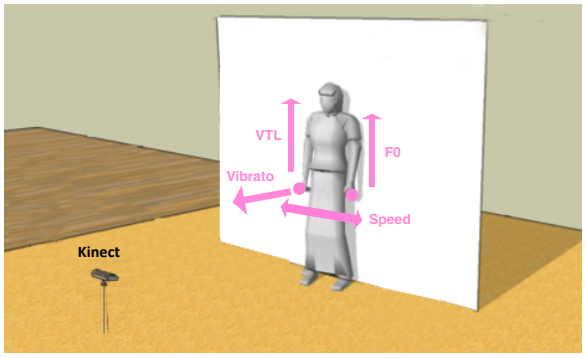

Figure 3: Mapping of the prosodic parameters.

This demonstration prototype has been displayed in various public spaces in Edinburgh as it provides a ludic approach to speech / singing synthesis for the general public.

4. Conclusions

We presented a demonstration prototype based on Mage/pHTS where different prosodic parameters of an HMM-based singing voice can be continuously controlled. This prototype has been designed originally to provide a ludic approach to speech synthesis for the general public but we believe that interactive speech synthesis has a large potential of useful applications. To mention one of them, we envision the use of interactive speech synthesis for voice-output communication aids that could generate spontaneous speech with minimal delay and allow the patient to modify the output speech on the fly.

5. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Koyabashi and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, in IE-ICE Transactions on Information and Systems, vol 83, no 11, pp. 2347-2350, 1999.

[2] T. Masuko, K. Tokuda, T. Koyabashi and S. Imai, “Voice characteristics conversion for HMM-based speech synthesis system”, vol 3, IEE, pp. 1611-1614, 1997.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Koyabashi and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system’, IEE, pp. 2523 -2526, 1997.

[4] H. Zen, K. Tokuda and A.W. Black, ‘Statistical parametric speech synthesis’, Speech Communication, vol 51, no 11, pp. 1039-1064, 2009.

[5] M. Astrinaki, N. d’Alessandro, B. Picart, T. Drugman, T. Dutoit. “Reactive and Continuous Control of HMM-based Speech Synthesis”. in Proc. Speech and Language Technology, Miami, USA, Dec. 2012.

[6] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System – Sinsy”, 7th Workshop on Speech synthesis, 2010, Japan.

[7] Z. Zhang, Microsoft Kinect Sensor and Its Effect, IEEE Multimedia Magazine, vol. 19, no. 2, pp. 4-10, 2012.

[8] “MAGE and HandSketch”, available at: http://vimeo.com/39558917”, March 2012.

[9] “MAGE and Face Tracking”, available at: http://vimeo.com/39567236”, March 2012.